Microstructural Feature Importance: Nowcasting Volatility

May 2020

Author: Kerem Proulx

Advisor: Andrew C. Spieler, Ph.D

Honors Essay, Department of Finance, Submitted in partial fulfillment of the requirements for the B.B.A. Degree, Department of Finance, Frank G. Zarb School of Business, Hofstra University, Hempstead, NY 11549-1340

Abstract:

Microstructural variables are used in the study of interaction between market participants and the underlying mechanisms of an exchange, as well between each other. These variables could be useful in detecting and measuring the probability of impending bouts of volatility when panicked behavior is observable. Using Financial Machine Learning techniques, this paper tests and studies the efficacy of several microstructural variables (The Roll Impact, Order-book Imbalance, V-PIN, Amihud's Lambda, and Kyle's Lambda) via the Mean Decrease Accuracy feature importance method trained on a Sequentially Bootstrapped Bagging Random Forest Classifier. Results show that V-PIN and Order Imbalance are most predictive when train using various labels of volatility (Standard Deviation of Returns, Corwin-Shultz Volatility, and Kurtosis).

Introduction

What moves prices? This fundamental question is one that permutates throughout finance and economics and is generally an underlying motivation for the development of various theories, models, and strategies. By nature, most of Finance is a zero-sum game. There is always someone on the other side of a transaction who suffers a loss due to your profit. A simple example is of a futures contract writer; by definition, if the contract buyer profits, the writer must subsequently realize a loss on that transaction. This is true for other classes of assets such as the bond market, such as when a firm issues callable bonds, they will likely call the bond when it is best for them and subsequently leaving the investor worse off (via reinvestment risk).

What many models and theories that rely on classical statistical methods and unrealistic assumptions of market friction fail to capture is the high-dimensionally complex nature of the interactions between market participants and the underlying mechanisms of the exchange. Under empirical market conditions, where price impact, transaction cost, and other such frictions affect the fundamental behavior of participants, it is important for researchers to understand the underlying structure of the markets they participate in. It is imperative then to codify these interactions in such a way that allows for measurement and the development of truly testable theories backed by empirical observations and thorough experimentation. The problem with adjudicating scientific disciplines to Finance, is that there is no true laboratory where theories can be tested, and variables isolated. This is especially true in modernity where electronification and interconnection has become the norm causing widespread multicollinearity and serial correlation of prices. Information is absorbed faster in diverse formats, affecting prices and behavior. Trades are executed at microsecond speeds; investment decisions are informed via satellite imagery and web-scraping techniques. Algorithms have become the most sophisticated and successful traders in the modern market. This paradigm, while posing many disruptive risks, presents some of the most interesting and complex problems of modern finance and economics.

As such, researchers must adopt new techniques and adapt to these changing market dynamics in order to remain competitive or risk being “outperformed” out of the business, or even fall prey to adverse selection and worst-case scenario: risk blowing up their accounts. Especially when they are affected by an event that the crowd perceives as a Black Swan and are then left scratching their heads as “bleeding-edge” firms continue to profit.

This paper serves as a study on the importance of modern financial variables pertaining to the underlying mechanisms and frictions of the markets (microstructure, as will be introduced in subsequent sections) and to assess their efficacy in predicting or measuring periods of large volatility. Volatility that may have been “un-predictable” by the masses but utilized as a source of alpha and opportunity by those able to detect its impending presence.

Section 1: Literature Review

1.1: Historical Background

1.1.1: High-Frequency Trading

Speed as a competitive advantage is not a new phenomenon in financial markets. It is said that in the 1800’s during the Battle of Waterloo, carrier pigeons of the House of Rothschild were used to inform Nathan Rothschild of the impending victory of the combined British and Prussian forces, enabling him to purchase distressed British bonds ahead of his competitors and capture a healthy profit [Kay, 2013].

High Frequency’s advantage arises in its usage of algorithmic and modern statistical solutions to the financial markets. As will be highlighted later in this paper, HFT does not view the market in the same paradigm as “Low-Frequency”. There are several aspects to this statement, but one such example is that HFT generally samples in event time rather than chronological. This allows for them to detect and measures various divergences from “expected” values or even to realize relationships in data that exhibits high-dimensionality and complexity that are not detectable by Low-Frequency: hierarchical relations, non-linear relations, etc.

HFT is not solely responsible for the generation of microstructural variables, as the study of market structure is a relatively old craft. However, HFT has allowed for the innovation and advancement of microstructure by innovating computational techniques that enable more effective measuring and calculation of market participant behavior. Perhaps it is true that HFT arose from the eventual tech-enabling of financial markets, however their persistence and search for a greater edge has led to some of the most disruptive and arguably some of the most valuable developments in modern finance (i.e. the overall increase in market liquidity due to the presence of HFT market makers and traders). HFT’s innate understanding of the inner-workings of the markets and the (albeit not always transparent) sharing of that information has led to a greater understanding of the financial markets today and has enabled many asset managers to adopt their techniques to better serve their clients.

1.1.2: Machine Learning vs Classical Statistics

The discipline of Statistics is a product of both its environment and time. Although this paper will focus on the efficacy and application of machine learning feature importance methods, it is important to understand the lineage of modern statistical methods and as they pertain to Finance.

Much of the statistical methods that are presented in university lecture halls today (i.e. linear algebra, regression analysis methods that rely on linearity, etc.) are a product of their time and are subject to computational restraints. Their popularity was likely due to their mathematical tractability, especially during the times of "by-hand" calculations and slide projectors [Efron and Hastie, 2016]. However, many of these methods are outdated and ill-suited to the modern (unstructured) datasets and problems that researchers and practitioners face today. Classical statistics was born during an era where the greatest computational power was the human mind. Today we have created machines that possess computational capacity that was only thought to be possible in stories of science fiction. As such, statistical methods have evolved, and Machine Learning has emerged as an invaluable tool to aid statisticians in making sense of unstructured, modern, granular and high-dimensionally complex data that has become the norm.

Machine Learning (ML) has been an area of scientific study under the broader engineering and computer science umbrella, for some time now. Beginning primarily with the perceptron developed by Frank Rosenblatt in the late 1950's. Although in its early days, ML was not the beast it is today, due to lack of technological and computational capacity; in modernity, Machine Learning has blossomed into one of the most "cutting-edge" technologies that has seen rapid adoption and caused mass disruption in industries from Farming to Consultation and has become somewhat of a buzzword in today's business vernacular, alongside "Big Data", Data Science, and the ever ominous "Artificial Intelligence".

The mysticism surrounding ML is generally attributable to its relative complexity and general hype and/or counter hype. Many believe ML to be a black box that produces crystal ball predictions with no real tractability or interpretability. Others see ML as a fantasy and claim it is forever prone to "overfitting". The issue with both ends of the cord is that ML, when used responsibly and correctly, is like any other of its classical statistical counterparts; it provides for a method of interpreting, understanding, and analyzing data to produce testable hypotheses. The notion that ML overfits or even underfits is not necessarily false, as, in the wrong hands ML will do more harm than good. However, much like any other statistical theory or practice, if used correctly and responsibly, Machine Learning is able to map, understand, interpret and explain unstructured data in an exceedingly high-dimensional space that exhibits great complexity and granularity and does so without the need for *a priori* model specification, human guidance and even without finite equation limits. The reason many see ML as a black box is a lack of understanding and perhaps due to many industrial applications placing preference on performance rather than theoretical understanding [Lopez de Prado, 2020a].

Finally, Machine Learning has begun to outpace classical statistical methods due to its high computational capacity. As such, ML algorithms and applications are able to "understand" and explain complex interactions within data that other statistical methods cannot, such as non-linear and hierarchical relations. ML can do so without the need for prior (ex-ante) knowledge of the correct specification of the problem being solved and without the need for unrealistic assumptions or parametric estimations. As will be discussed later, ML disentangles the variable search from the specification search [Lopez de Prado, 2018] leading to greater robustness than classical methods. Perhaps, the reader should observe classical statistics as a precursor to Machine Learning, much like how weight-lifters must first learn proper technique and form before moving to more advanced weight-classes and complex lifts that will serve to increase their strength and muscle mass.

1.1.3: Financial Machine Learning as a Discipline

As it pertains to Finance, Machine Learning (ML) is not a one-size-fits-all application. Applying ML to financial datasets is much different than teaching an algorithm to play chess. Financial and economic systems pose a degree of complexity that is beyond the scope of classical statistical methods [Lopez de Prado, 2018]. As it pertains to finance, ML is an exceptional tool in an asset manager's arsenal. However, before applying ML techniques to financial datasets, a practitioner and/or researcher must first understand the nature of the task they wish to accomplish or the problem they wish to solve. It is not as simple as training an off-the-shelf model on some data and being able to predict price movements. In fact, price prediction is only one of the myriads of applications available to Financial ML researchers. Financial datasets exhibit extremely low signal-to-noise ratios, and as such require specialized techniques to treat the data before any algorithm can extract value. Financial signal also decays rapidly over time as trades become crowded. This adds an extra layer of complexity as models must be deployed dynamically to solve constantly changing problems. Financial datasets also exhibit attributes such as serial-autocorrelation and non-linear substitution effects that create several roadblocks for researchers to address before any model or algorithm can be applied.

Financial ML is not meant to be used as a crystal ball approach to the markets. No algorithm will tell you exactly how and what bets to place. If that were the case, this paper would not exist, and all alpha would be arbitraged out of the system. Where financial ML shines is in its ability to aid in theory-crafting. Of course, as mentioned, financial ML has a myriad of applications ranging from bet-sizing, feature importance, price prediction, clustering, portfolio optimization, detection of false discovery, etc.; it is most useful when applied holistically and as a tool to generate a testable theory/hypothesis to then develop an investment strategy/"trading rule" [Lopez de Prado, 2020a].

1.2: Unstructured Data and the Folly of p-values

According to SINTEF (2013), 90% of the world's data has been generated in the "last two years" (the accuracy of this statement may be slightly skewed as this paper had been released in 2013, but nonetheless emphasizes the point). The IDC (2014) found that 80% of the worlds data is unstructured. In an era that requires greater sophistication and complexity in the statistical practice, feature Importance Analysis is a necessary research tool in any investor's arsenal. It is used to determine the significance or predictive power of particular variables. This is especially useful in finance as researchers are provided greatly unstructured datasets that exhibit high-dimensionality, sometimes with the number of variables even exceeding that of the number of observations/samples. This new paradigm generally renders the use of classical methods, such as p-values, moot. P-values suffer multiple flaws which will be discussed further in this paper. So much so that the president of the American Finance Association, Campbell Harvey, has acknowledged that most discoveries in finance are false [Harvey, 2017].

Nuzzo [2014] describes the pitfalls and origination of the p-value as measure of statistical significance in broader scientific applications. According to the paper, p-values were developed by Ronald Fisher, a classical statistician, as a method of testing whether findings/evidence was significant enough for a deeper dive. P-values were meant to distinguish whether experimental results were consistent (or not) with what random noise could produce. A somewhat irrelevant probability as discussed later.

P-values are widely known and taught classical statistical tool. However, as stated above, they suffer from being ill-prepared to handle high-dimensionality and complex interactions of data. Lopez de Prado (2020b) and Colquhoun (2014) defines the pitfalls of p-values as follows: 1) p-values are unable to disentangle the Specification search from the variable/significance search. This is due to the assumption that, as p-values are applied to estimated coefficients of models, that a given model is correctly specified. In financial systems where the complexity only allows for rough guessing of model specification, p-values will not be able to capture complex interactions, and likely lead to false discoveries. 2) In-sample estimation errors are used when computing p-values, especially via Ordinary Least Squares methods that attempt to estimate variance in-sample (i.e. performance on train data). The issue with this approach is that information leakage occurs as each data point used to train the model is also used to test the model. This is a cardinal sin in Machine Learning and will most definitely lead to model overfitting and poor performance on unseen data (out-of-sample). As such, ML researchers typically weight the generalization error with a higher importance. 3) P-values compute a rather useless probability. The chance that there is no real effect in the results. What really should be tested is whether there is a real effect caused by random variable. This attribute leads to false discovery rates being exceedingly higher than the classical theory of 5%. According to Colquhoun [2014], in a *perfect* experiment the false discovery rate is upwards of 30% when using a p-value of 0.05. In Finance where the signal-to-noise ratio is low and false positive rate is naturally high, false discovery rates at a p-value of 0.05 are exceedingly higher. 4) A large issue in the scientific community and especially in the financial community, is selection bias under multiple testing.

Researchers seldom publish the amount of trials and the details of those trials when they publish results, instead they select the model or back test that garnered the best results (lowest p-value). The issue here is that, when performing multiple experiments on the same dataset, the false positive rate exponentially grows in relation to number of trials ($\alpha[K] = 1 - (1-\alpha[1])^K$; where K = the number of trials). This leads to exceedingly high false discovery rates, which is likely why many published "strategies" produce irreproducible results and fail when implemented. 5) Finally, p-values rely on unrealistic assumptions of the underlying data, especially in modernity. They require a lack of multicollinearity, white noise residuals, normally distributed residuals, and are not robust to outliers. As such, they may assign abnormally large p-values to predictive variables, falsely dismissing them as irrelevant.

This paper utilizes a well-known and tested feature importance method; Mean Decrease Accuracy (MDA). This method overcomes many of the flaws of p-values and offer more robust findings of importance/predictive power of variables, providing a more stable and efficient solution to the search for variables/parameters. MDA well researched in Machine Learning literature; Park, Hur and Ihm (2017) and Han, Guo and Yu (2016) demonstrate the use of the Mean Decrease Accuracy method as a means of assessing variable impact to Random Forest decision-making as applied to the mobile cloud computing and software engineering industries with the goal of demonstrating model interpretability; opening the proverbial Black-Box of ML.

1.2.1: Mean Decrease Accuracy: A Machine Learning Solution to p-values

Lopez de Prado (2020a) establishes the MDA method as it pertains to financial ML. At its core, MDA is a feature analysis method used to determine variable impact (predictive power) on a model's decision-making (predictions). The most common process of applying MDA is via Cross-Validation scoring and feature shuffling. (1) Fit a model on a set of features (variables), irrespective of model specification (MDI is a random-forest specific method, however). In this paper we will be utilizing a random-forest algorithm. (2) Compute and cross-validate the prediction scores. Lopez de Prado (2018) specifies the Purged K-fold Cross Validation method. (3) Iterate through every feature, shuffling the feature and refitting the model on the newly shuffled feature and cross-validate the new prediction score. (4) Compute the distribution of the generalization error between the CV scores of steps (3) and (2) to determine the loss in model performance due to shuffling. If a shuffle causes a large loss in model performance, the importance of the feature shuffled can be measured. Lopez de Prado (2019b) determines that MDA can utilize various scoring methods and is not restrained to accuracy as a performance measure, making this method highly adaptive and robust to various problems.

Lopez de Prado (2018) demonstrates MDA's ability to avoid selection bias that is prevalent in finance and overfitting that can occur with irresponsible use of ML. He describes the Combinatorial Purged K-Fold Cross-Validation (CPCV) method to be used when determining feature importance. CPCV is a leave-p-out Cross Validation process and is used to prevent overfitting to both test and train sets and to prevent selecting a model based on performance of a particular split (selection bias under multiple testing).

The number of possible splits formed from a CPCV process is defined by Lopez de Prado (2018) as:

$$\binom{N}{N-k} = \frac{\prod_{i=0}^{k-1}(N-i)}{k!}$$

Where *N* is the number of T observations partitioned into groups and *k* is the size of each group. As each combination includes *k* tested groups, the total groups tested is defined as:

$$k\binom{N}{N-k}$$

Since all possible combinations are computed, each test group is uniformly distributed across all *N* as each group belongs to the same number of train and test sets.

The total number of paths generated for testing, implying k-sized sets over N groups ($\varphi[\mathrm{N},\mathrm{k}]$), is defined as:

$$\varphi[\mathrm{N},\mathrm{k}] = \frac{k}{N}\binom{N}{N-k} = \frac{\prod_{i=1}^{k-1}(N-i)}{(k-1)!}$$

For instance, with N=6 and k=2, we derive 15 possible splits, and 5 paths generated. These combinations scale exponentially with greater values of N and k, requiring greater computational power. For example, with N = 8 and k =4, we can derive 70 possible splits formed. With 35 sample paths.

Lopez de Prado (2018) also asserts that the train sets should then be purged and embargoed to ensure no test-data leakage.

1.2.3: Selection Bias

Harvey et al. (2016) and Harvey (2017) ascertains that the current problem of selection bias, overfitting, the publishing of false discoveries and p-hacking was not caused by ML but have been symptoms of poor econometric and outdated statistical practices in financial research.

The issue of multiple testing and selection bias has been studied and examined throughout the statistical literature for years. See Bailey et al. (2014), Benjamini and Liu (1999), Benjamini and

Hochberg (1995) and Lopez de Prado (2018). As such, why is it that most financial researchers and the majority of the literature fail to recognize the effects of such issues?

Romer (2016) and Solow (2010), criticize the use of classical Economic theories as “facts with unknown truth values” and “generally phony assumptions”. These beliefs are generally due to the reliance on unrealistic assumptions, un-robust estimation of parameter values, misspecification, overfitting, and lack of complexity and generalization that classical econometric models exhibit.

Lopez de Prado (2020a) provides basis on what makes ML feature importance methods robust is their ability to disentangle the variable search from the specification search. Further ascertaining, that ML’s role is not that of an oracle, but to assist the researcher in developing theories that can explain high-dimensional complexities. Once theories or relationships are identified they can then be tested for correlation, codependence or causation as examined by Pearl (2009) and Wright (1921).

1.3: Microstructure

O’Hara (1998), O’Hara (2014) and Easley et al. (2012) define Microstructural Theory as the study of the underlying, granular processes that affect transactions and executions under explicit market structures or trading rules. In short, Microstructural features are aimed at discerning information from trade data of market participants to examine how they both conduct their trading and attempt to conceal their intentions from other participants. Generally, institutions would be able to access FIX protocol messages via purchasing data streams from exchanges and use this data to discern intentions, patterns and build informative features. However, microstructural theory has developed features and methods of extracting trade information from market data, following theories of asymmetric information and subsequent adverse selection of market makers by informed traders. Testing for feature importance of these variables allows researchers to formulate theories that can then be tested via looking for evidence of market participant behavior within FIX messages (i.e. whether market makers become liquidity takers rather than providers during times of market panic).

There are several such variables that will tested in this paper.

1.3.1: Roll Impact

The Roll model [Roll, 1984] was one of the initial “first generation” models developed to explain the effective bid-ask spread of an instrument. Its primary function was as a proxy for measuring liquidity of an instrument as bid-ask spread are generally determined as a function of liquidity. The tighter the spread, the greater the liquidity. It is used primarily in securities that are thinly traded or where the published quotes are not indicative of actual levels at which market makers are willing to provide liquidity [Lopez de Prado, 2018].

Consider a Random Walk with no drift to represent a series of mid prices: *{m_t}*

$$m_t = m_{t-1} + u_t$$

Changes in price are then assumed to be drawn from an IID normal distribution:

$$\Delta m_t \sim N[0, \sigma_u^2]$$

This assumption quite obviously breakdowns in real market conditions and defies empirical observation as financial price series are generally do exhibit a drift, heteroskedasticity, serial dependency and of course their returns are non-Normal.

However, if sampled properly, as will be discussed in later sections, the Roll model determines that prices are a result of trading against the bid-ask spread [Lopez de Prado, 2018].

$$p_t = m_t + b_t c$$

Such that, c = $\frac{1}{2}$ b_t is the aggressor side defined by the tick rule defined in the data section. Assuming buys and sells are equally as likely; P[b_t = 1] = P[b_t = -1] = $\frac{1}{2}$, are serially independent, and independent of noise; the values of c and σ_u^2 are derived as:

$$\sigma^2[\Delta p_t] = E[(\Delta p_t)^2] - (E[(\Delta p_t)])^2 = 2c^2 + \sigma_u^2$$

$$\sigma[\Delta p_t, \Delta p_{t-1}] = -c^2$$

Such that:

$$c = \sqrt{max\{0, -[\Delta p_t, \Delta p_{t-1}]\}}$$

And

$$\sigma_u^2 = \sigma^2[\Delta p_t] + 2\sigma[\Delta p_t, \Delta p_{t-1}]$$

This describes the bid-ask spread as a function of serial covariance in price changes and that unobserved noise is a function of the observed noise and the serial covariance of price changes [Lopez de Prado, 2018].

1.3.2: Order Flow Imbalance

Shen (2015) establishes that Order Imbalance generally exhibits a positive relationship with returns as traders post limit orders to either buy or sell they have an impact on the volume of bid-ask on the limit-order book; therefore revealing their intentions. Insight into these intentions allow for information leakage on the t+1 price movement.

Utilizing the tick rule defined in the below section (assigning an aggressor of either buy {1} or sell {-1} to each trade in the order book) and sampling by Volume Bars, also defined below, we are able to establish the Bulk Volume of buy-aggressor trades:

$$C_{b,t} = b_t^+ V_t$$

Where $C_{b,t}$ is the Cumulative buy volume for each bar *t*, b_t^+ is the total number of buy aggressor flags within bar *t* and V_t is the total volume of bar *t*.

Imbalance can then be computed as:

$$\theta = \left(\left(\frac{C_{b,t}}{V_t} \right) * 100 \right) - 50$$

Since imbalance is computed as a function of bulk buy volume, it is measured as a divergence from a halfway point (50) that is considered sufficiently "balanced" (i.e. relatively equal volume on the bid and ask) and any divergence from the halfway point signals an imbalance in the limit order book. Order imbalance is a well-known proxy for discerning the presence of informed traders as the persistence of imbalance may produce a signal as to the intent of the trade. Toth et al (2011) studies the effects of persistent order imbalance on market impact and how order imbalance autocorrelation is a product of order splitting of informed/large traders.

1.3.3: V-PIN

A key measure of the probability of adverse selection by informed traders is the VPIN method developed by Easley et al. (2012a) as a high-frequency paradigm evolution of the PIN method developed by Easley et al. (1996). The Volume-Synchronized Probability of Informed Trading is as follows:

Take a security with price S. Its price at time t=0 is denoted S_0. The expected price of the security at time *t* can be calculated as:

$$E[S_t] = (1 - \alpha_t)S_0 + \alpha_t[\delta_t S_B + (1 - \delta_t)S_G]$$

Where, α is the probability that new information will arrive within the timeframe, δ_t is the probability that news will be bad (subsequently $(1-\delta_t)$ is the probability of good news) and S_B and S_G is the new price of the security following the incorporation of new information (either Bad or Good).

Assuming a Poisson distribution, informed traders enter the order book at rate μ, and conversely uninformed traders arrive at rate ε. Therefore, to maintain a neutral position and void loss, market makers must establish a breakeven bid level of B_t, defined as:

$$E[B_t] = E[S_t] - \frac{\mu\alpha_t\delta_t}{\varepsilon + \mu\alpha_t\delta_t}(E[S_t] - S_B)$$

The breakeven ask is defined as, A_t:

$$E[A_t] = E[S_t] + \frac{\mu\alpha_t(1-\delta_t)}{\varepsilon + \mu\alpha_t(1-\delta_t)}(S_G - E[S_t])$$

The breakeven spread is then calculated as:

$$E[A_t - B_t] = \frac{\mu\alpha_t(1-\delta_t)}{\varepsilon + \mu\alpha_t(1-\delta_t)}(S_G - E[S_t]) + \frac{\mu\alpha_t\delta_t}{\varepsilon + \mu\alpha_t\delta_t}(E[S_t] - S_B)$$

Assuming a base case of $\delta_t = ½$ (i.e. 50% probability of bad news):

$$\delta_t = \frac{1}{2} \Rightarrow E[A_t - B_t] = \frac{\alpha_t\mu}{\alpha_t\mu + 2\mathrm{ε}}(S_G - S_B)$$

Therefore, the critical factor in determining the level of the bid-ask spread where market makers will provide liquidity is:

$$PIN_t = \frac{\alpha_t\mu}{\alpha_t\mu + 2\varepsilon}$$

To compute PIN, the estimation of four non-observable factors is required: {α, δ, μ, ε}, via a maximum likelihood model to fit a mixture of three Poisson distributions and is defined as:

$$P[V^B, V^S] = (1-\alpha)P[V^B,\varepsilon]P[V^S,\varepsilon] + \alpha(\delta P[V^S,\varepsilon]P[V^B,\mu+\varepsilon] + (1-\delta)P[V^B,\mu+\varepsilon]P[V^S,\varepsilon])$$

V^{B} is the volume of buy-aggressor initiated trades (against the ask) and V^{S} is subsequently the volume of sell-aggressor initiated trades (against the bid).

Easley et al. (2008) formulated that:

$$E[V^B - V^S] = (1-\alpha)(\varepsilon - \varepsilon) + \alpha(1-\delta)\big(\varepsilon - (\mu + \varepsilon)\big) + \alpha\delta(\mu + \varepsilon - \varepsilon) = \alpha\mu(1-2\delta)$$

And for an exceedingly large μ;

$$E[|V^B - V^S|] \approx \ \alpha\mu$$

The Volume-Synchronized Probability of Informed Trading is a high frequency paradigm of PIN and synchronizes data sampling with market activity (volume-time), aligning with high-frequency sampling methods. Therefore, it is estimated that;

$$\frac{1}{n}\sum_{\tau=1}^{n} |V_\tau^B - V_\tau^S| \approx \alpha\mu$$

Where V_τ^B and V_τ^S are the sum of volumes of either buy-aggressor or sell-aggressor trades in volume bar τ, while n is the number of bars used to produce the estimate. As all volume bars are of the same size, it can then be formulated that:

$$\frac{1}{n}\sum_{\tau=1}^{n} (V_\tau^B + V_\tau^S) = V = \alpha\mu + 2\varepsilon$$

As such, the high-frequency paradigm of PIN is defined as:

$$VPIN_\tau = \frac{\sum_{\tau=1}^{n}|V_\tau^B - V_\tau^S|}{\sum_{\tau=1}^{n}(V_\tau^B + V_\tau^S)} = \frac{\sum_{\tau=1}^{n}|V_\tau^B - V_\tau^S|}{nV}$$

1.3.4: Amihud’s Lambda

Amihud’s Lambda [Amihud, 2002] finds a positive relation between absolute returns and illiquidity by computing the daily price response associate with one dollar of trading volume and arguing that its value is a strong proxy of price impact. This is concurrent with the theory that sufficiently liquid markets should be able to absorb large volumes with minimal price impact. Lopez de Prado (2018) implements Amihud’s Lambda as:

$$|\Delta \log\,[\tilde{p}_\tau]| = \lambda \sum_{t \in B_\tau} (p_t V_t) + \varepsilon_\tau$$

Where B_τ is the set of trades within bar τ, p_t is the closing price of bar τ, and V_t is the dollar volume in trade $t\epsilon B_\tau$; λ captures the price impact.

1.3.5: Kyles Lambda

Kyle (1985) defines a trade model that consists of:

$$v{\sim}N[p_0, \Sigma_0]$$

Where *v* is the terminal value of a "risky" asset. Further, two traders are considered; one noise trade that transacts quantity $u = N[0, \sigma_u^2]$ that is independent of *v*; and an informed trader who knows *v* and orders quantity *x*.

A market maker will experience order flow as a function of y = x + u and will adjust the price *p* to as a result. It is assumed that market makers are not able to distinguish from informed or noise traders. Order flow imbalance as defined above may be indicative of the presence of informed traders, forcing the market maker to adjust prices. As such, a positive relationship between change-in-price and flow imbalance (i.e. market impact of orders).

An informed trader assumes the market maker exhibits a linear price adjustment function of $p = \lambda y + \mu$ with λ representing the inverse measure of liquidity (the higher lambda value the less liquidity).

The trader's profits are formulated as $\pi = (v - p)x$, maximized at $x = \frac{v-\mu}{2\lambda}$ conditional on λ > 0.

We can then assume the market maker formulates the trader's probable demand as a linear function of *v* where $x = \alpha + \beta v$; implying $\alpha = \frac{\mu}{2\lambda}$ and $\beta = \frac{1}{2\lambda}$.

Kyle then stipulates that the market maker finds equilibrium level between maximum profit and maximum efficiency at:

$$\mu = p_0$$

$$\alpha = p_0\sqrt{\frac{\sigma_u^2}{\Sigma_0}}$$

$$\lambda = \frac{1}{2}\sqrt{\frac{\Sigma_0}{\sigma_u^2}}$$

$$\beta = \sqrt{\frac{\sigma_u^2}{\Sigma_0}}$$

The informed trader's expected profit can then be defined as:

$$E[\pi] = \frac{(v - p_0)^2}{2}\sqrt{\frac{\sigma_u^2}{\Sigma_0}} = \frac{1}{4\lambda}(v - p_0)^2$$

The informed trader's sources of profit are then defined as: the security's mispricing, the variance of the noise trader's order flow (with greater noise, the informed trader can mask their intentions/footprint) and the reciprocal of the security's variance (low volatility, the easier monetization of mispricing).

As such, λ captures overall price impact, with illiquidity raising in tandem with uncertainty of terminal value, *v,* and falls with greater noise. In practice, Kyle's Lambda is estimated via regression:

$$\Delta p_t = \lambda(b_t V_t) + \varepsilon_t$$

Wherein; $\{p_t\}$ is a series of prices, $\{b_t\}$ is a series of aggressor side flags $\epsilon\{-1,1\}$ and $\{V_t\}$ is a series of trade volume. Hence, (b_tV_t) is the computation of signed net-order flow.

Section 2: Data & Methodology

2.1: Trade Level Tick Data

Tick level data of the SPDR S&P 500 ETF Trust gathered from Polygon.io ranging from April 8th of 2017 to April 8th of 2020 was sampled. A tick is generated every time a trade occurs and is recorded via the exchange's matching engine network protocol, that output FIX (Financial Information eXchange) messages containing trade information. These datasets are generally quite large as there are millions of data points/trades spanning a given historic time frame; for instance, the raw tick data generated for this study amounted to about 3gbs of total size and over 37,000,000 observations. Handling of data this large generally requires ample computing power and as such a virtual Google Cloud instance was utilized with 960gb of RAM and 40 CPU cores (even with such raw compute power, processing and fitting the data took hours at a time) with the use of multithreading and other HPC techniques.

SPY raw tick data is characterized by a timestamp measured in microseconds, a unique trade identifier, the size of each trade, the price at which the trade was executed, a unique exchange identifier and any condition flags. The SPY ETF was chosen primarily due to the low cost of acquiring the tick level data but in addition to it being the oldest and generally most liquid and widely traded ETF on the market.

The data was processed first by decomposing order flow via the Tick Rule (defined below) and generating trade bars sampled by a pre-determined threshold of volume buckets (i.e. whenever a threshold of volume is reached, a bar is sampled consisting of all trades running-up to that bar; this is consistent with the event time sampling techniques employed by HFT [Easley et al., 2012b]), then generating the various microstructural variables defined in the above section through different estimation windows (50, 100, and 250 bars).

The tick rule defines a sequence of labels, $\{b_t\}_{t=1,\ldots T}$ generated from a tick sequence where:

$$b_t = \begin{cases} b_{t-1}\, if\ \Delta p_t = 0 \\ \frac{|\Delta p_t|}{\Delta p_t} if \Delta p_t \neq 0 \end{cases}$$

The result is a sequence that represents an estimation of signed order-flow where $b_t \in \{-1,1\}$.

Volume Synchronized bars were generated using a lower bound dynamic volume threshold of 28,000 per bar where each bar's total volume/size ranges from 28,000 to 30,000; 267,159 bars were sampled in total. This sampling technique allows for a partial return to normality of returns as depicted in figure A, making for much more desirable properties when conducting analyses. A bar is sampled whenever aggregate tick volume crosses the threshold; each bar is made up all ticks running up to said bar. This is done iteratively until all ticks have been binned into bars.

2.2: Target Values

2.2.1: Standard Deviation of Bar-to-Bar Returns

Bar-to-bar returns are then calculated via a simple percent change/holding period return formula where the time period is a single bar and the only cash flow being the change in price from bar to bar. The standard deviation of bar-to-bar returns is used as a proxy measurement for volatility. This is calculated on a rolling window of 100 bars throughout the study as a constant basis for measurement. This method is consistent with the theory that a sufficiently liquid market (i.e. a market where market makers are liquidity providers not takers) that is of large size (i.e. the market for SPY shares) should generally experience low standard deviations of returns during times of low volatility and conversely may experience higher standard deviations of returns during periods of high volatility which may be caused by sufficient illiquidity due to the interactions between trader behavior and market maker behavior (i.e. marker makers becoming liquidity takers)consistent with the general theory of liquidity that a sufficiently liquid market should have the ability to absorb large order flow with minimal price impact. This can be observed in figure B; plotting the standard deviation of bar-to-bar returns from 2017 to 2020. A sharp increase in standard deviations can be observed during the period of 2020-02 to 2020-04 attributable to the market panic caused by the COVID-19 crisis wherein markets experienced widespread sell-off and large impacts on absolute returns over several days. This serves as an effective testing period for experimentation.

2.2.2: Corwin-Shultz Estimator

Corwin and Shultz (2012) formulated a bid-ask estimator as a function of high-low prices. There are two principles captured in this model: high prices are generally always matched against the offer and lows are generally always matched against the bid. The ratio between high-low prices represents a measure of volatility while also reflecting the bid-ask spread. The second principle of the model holds that the high-low ratio component that captures volatility, proportionately increases with time between subsequent observations.

Spread can be computed as:

$$S_t = \frac{2(e^{\alpha_t} - 1)}{1 + e^{\alpha_t}}$$

Such that,

$$\alpha_t = \frac{\sqrt{2\beta_t} - \sqrt{\beta_t}}{3 - 2\sqrt{2}} - \sqrt{\frac{\gamma_t}{3 - 2\sqrt{2}}}$$

$$\beta_t = E\left[\sum_{j=0}^{1}\left[\log\left(\frac{H_{t-j}}{L_{t-j}}\right)\right]^2\right]$$

$$\gamma_t = \left[\log\left(\frac{H_{t-1,t}}{L_{t-1,t}}\right)\right]^2$$

$H_{t-1,t}$ is the high price over 2 bars (prior and current) and $L_{t-1,t}$ is the low price over the same interval. Corwin and Shultz recommend setting all negative alpha values to zero as $\alpha_t < 0 \Rightarrow S_t < 0$.

2.2.3: Kurtosis

Excess Kurtosis of bar to bar returns is used to measure the efficacy of microstructural variables in detecting heavy tails. As such the Excess Kurtosis formula is defined as:

$$Kurt = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^4/N}{s^4} - 3$$

Computed over a rolling window of 100 bars.

2.3: Down-sampling via Filters (CUSUM Filter)

Observations within the series of standard deviations are filtered using a CUSUM filter that detects a shift in the mean value of a series from a target range. Bar *t* is sampled if and only if the cumulative sum of the series exceeds or equals a threshold *h*. The cumulative sum is then reset once the threshold is repeat. Lopez de Prado (2018) defines the filter as:

$$S_t = max\{0, S_{t-1} + y_t - E_{t-1}[y_t]\}$$

with boundary condition of $S_0 = 0$, where; $\{y_t\}_{t=1,...,T}$ is a series of observations sampled from a IID, locally stationary process.

In this study a CUSUM filter of size 0.05 was used to sample observations of a series of standard deviations. This resulted in a down-sampled size of 18,275 filtered observations. Figure C plots a Geometric Brownian Motion with sampled observations shown in red using a filter size of 0.05.

2.4: Labelling (Trend-Scanning)

Label generation is a particularly important process in machine learning as it determines the clear specification of the problem to be solved. For this study, we will assign labels to observations of standard deviations (volatility) consisting of $L_t \epsilon$ {-1,0,1} such that an observation within a downtrend is assigned a label of -1 and an observation within an uptrend is assigned a label 1 and an observation within no clear trend is assigned the label 0. Trends are determined using the Trend Scanning Method [Lopez de Prado, 2020] developed by Marcos Lopez de Prado, Lee Cohn, Michael Lock and Yaxiong Zeng. First, a t-value is computed that is associated with the estimated regressor coefficient in a linear time-trend model with a set look-forward period. Issues arise when various look-forward periods produce varying t-values. As such, the researchers determined a method of scanning different look-forward period for the value that maximizes the t-value, subsequently assigning the most statistically significant trend to the observation. T-values are then used as sample weights when training our classifier. Figure *D* depicts an example of the trend scanning method applied to a Gaussian Random Walk with a trend where t-values are used to adjust the hue of observations to show the significance of trend.

Section 3: Modelling

3.1: Sequentially Bootstrapped Bagging Classifier

This study utilizes a Sequentially Bootstrapped Bagging Classifier that fits a base Random Forest Classifier of 1,000 estimators on random subsets of the data using a Sequential Bootstrapped sampling method. The predictions are then aggregated via a voting method (a large probability of an observation being within a "class" is assigned when a higher proportion of estimators classify said observation accordingly) to form the final prediction. This is done to address the well-known variance issue that occurs when predicting using a fit Decision Tree or Random Forest Classifier and works by generating N random datasets (sampled from the current dataset) with replacement, then fitting an equal number of estimators, one for each dataset sampled. The predictions are then aggregated as mentioned above. Bootstrap Aggregation (Bagging) has been well studied in Machine Learning literature to reduce a models overall forecast variation as a direct function of the number of estimators, average variance of predictions and the correlation among predictions while also increasing accuracy of the Bagged meta-Classifier [Lopez de Prado, 2018].

Feature importance is then calculated via the Mean Decrease Accuracy method defined in the prior section, utilizing the out-of-bag score derived from fitting the Sequentially Bootstrapped Bagging Classifier and an out-of-sample score derived from the Purged K-Fold Combinatorial Cross Validation defined in the prior section. The Feature importance is then plotted and discussed in the results section.

Section 4: Results

Figures F – H depict MDA graphs displaying the feature importance determined by the Sequentially Bootstrapped Bagging Classifier trained on a subset of data. It is important to note that as MDA is computed on out of sample performance, features may be deemed completely un-important or even detrimental to model performance. The value in this function is that researchers are able to identify features that can be removed from the feature matrix in order to limit model complexity and provide for more efficient use of computational resources (as model train time will be reduced as a result of a smaller input matrix). Various estimation windows were used to determine feature importance; 50, 100 and 250 bars.

Each figure plots the feature importance in blue and maps the standard deviation of importance over the estimation period in red. Out of bag and out of sample scores are reported for each test as well. Model performance generally converged around .57 (57%) accuracy for out of bag estimation and about .51 (51%) for out of sample estimation when predicting Standard Deviation of bar to bar returns. This is relatively good performance for a financial ML classification model.

However, performance converged to about 74% accuracy out of bag and 78% out of sample when predicting the Corwin-Shultz Volatility measure. This accuracy could be improved by removing a unpredictive features over the 250-bar window (Kyle's Lambda and VPIN), in favor of Order Imbalance. Imbalance is likely a strong predictor of the Corwin-Shultz measure as there is strong evidence in the literature that Order Flow Imbalance that is highly persistent has a large affect on the movement of bid-ask spreads set by market makers as they attempt to maintain their neutral positioning and continue to provide liquidity.

Interestingly, over the 50 bar estimation window only Kyle's Lambda was able to somewhat predict the heavy tailed-ness of returns, whereas over longer estimation periods, none of the microstructural variables were able to detect excess Kurtosis, and actually hindered model performance. This suggests that there are other factors at play as it pertains to the empirical detection of heavy tails.

Section 5: Conclusion

Microstructural variables have been utilized by quantitative traders ever since mathematics and computational and scientific disciplines began to apply their practices to the financial markets. High-dimensional complexities require complex algorithms, models and variables to understand the infinitely complex interactions between data. Microstructural variables create a framework that enables Machine Learning to adequately measure and understand these complex interactions and aid researchers in crafting theories that may serve to explain certain phenomenon experienced in the markets. Are these techniques necessarily restricted to High-Frequency traders and Physics PHDs that are seen as some sort of alchemists with crystal ball-like forecasts? Hopefully this paper shows that even lower-frequency traders and more "traditional" assets managers and researchers can leverage modern statistical and machine learning techniques to discover a deeper understanding of the mechanisms "under the hood" of the markets and make more informed investment decisions.

5.1: Microstructure and Nowcasting: Detecting Earthquakes

The author would be remised if he did not mention the importance of nowcasting and variable measurement as they pertain to finance. There is a stark difference between creating models that produce long-term forecasts (and subsequently are subject to greater risk of estimation error) and models that, via nowcasting and measurement, produce an actionable short-term signal that may indicate the presence of certain phenomenon.

Nowcasting is utilized in a broad range of complex sciences, but one of the most descriptive examples is that of Earthquake Detection. Researchers are not able to accurately predict the occurrence of earthquakes by any means. They can however measure certain variables and conditions that are generally indicative of an earthquake, allowing them ample time to warn localities to evacuate their population, sometimes with even hours' notice.

This practice allows for perceived Black Swans problems to become easily detectable White Swans. Take the widespread market selloff related to COVID-19 that occurred in late February into early March. Many investors, traders, large institutional investors, retirement plans, etc. suffered huge losses due to the toxically volatile drop in prices. However, many Market Making firms were relatively un-harmed. Why is this? Due to their knowledge of the inner workings of markets, they were able to nowcast (measure) order flow imbalances, toxic order flow and other such variables to inform their defensive positioning. Afterall, Market Makers are participants in the markets and aim to hold "neutral" positions to capture the profit implicit in the bid-ask spread. Of course they would concern themselves with the measuring of market conditions to protect themselves from contagion.

Figure E plots the COMEX High Grade Copper Cash future prices (in orange) and the S&P 500 Index (in black). On January 23rd, the Chinese government officially imposed strict lockdown measures on afflicted provinces, causing a huge shock to manufacturing activity. This caused a selloff in commodities linked to industry. Additionally, those who utilized alternative data sources were able to measure the impact on the Chinese supply chain virus had, signaling adverse market environments ahead. The selloff in the US market ensued only weeks after, ending the longest bull-market run in history. Who survived? Those who nowcasted and bought protection early.

The importance of microstructure in modern finance inherent in its ability to capture high-dimensionally complex, non-linear interactions between market participants and the information their actions convey. To continue to maintain an edge and generate value for their clients, asset managers should begin to adopt more modern statistical and computational methods such that they are able to leverage nowcasting and develop theories through which they can capitalize on hard-to-find market phenomena. Although microstructure has been studied for decades, the combination of modern computational capacity and machine learning allow for asset managers to properly detect and protect themselves from toxic contagions that exist in the markets today.

Figure A

Partial Return to Normality via Volume-Synchronized Bar-based Sampling

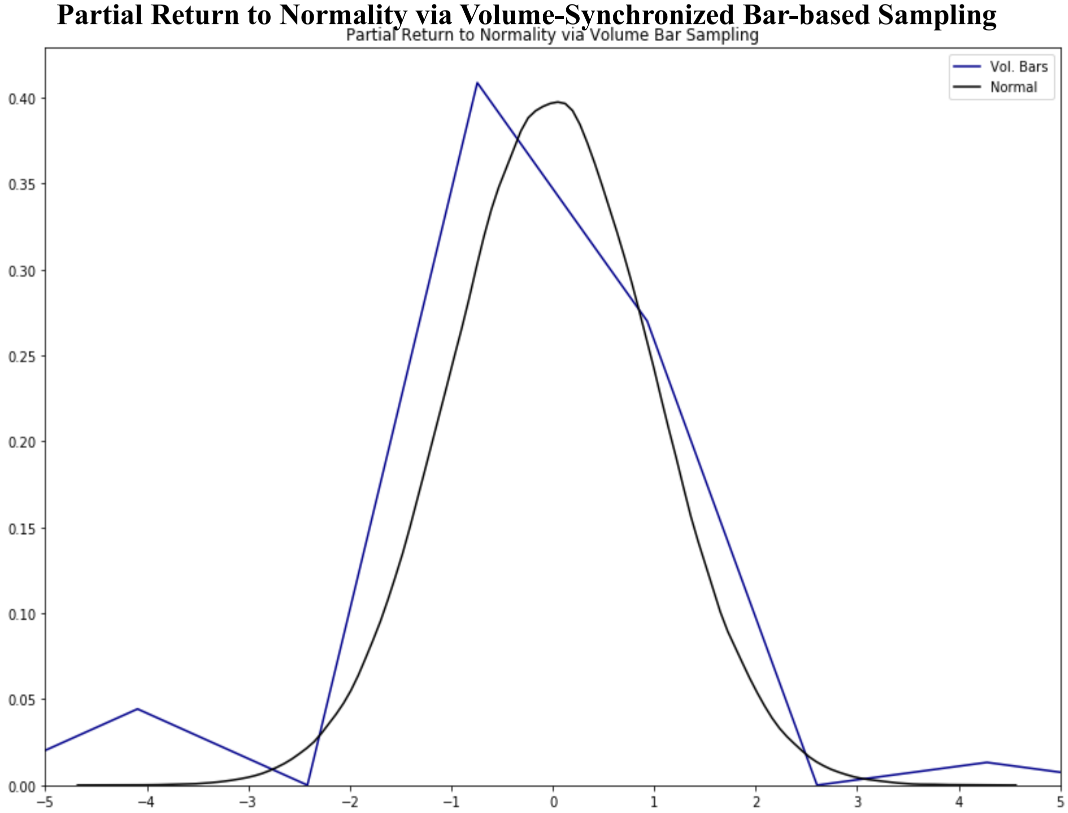


Figure B

Standard Deviation of Bar-to-Bar Returns Over Sampling Period

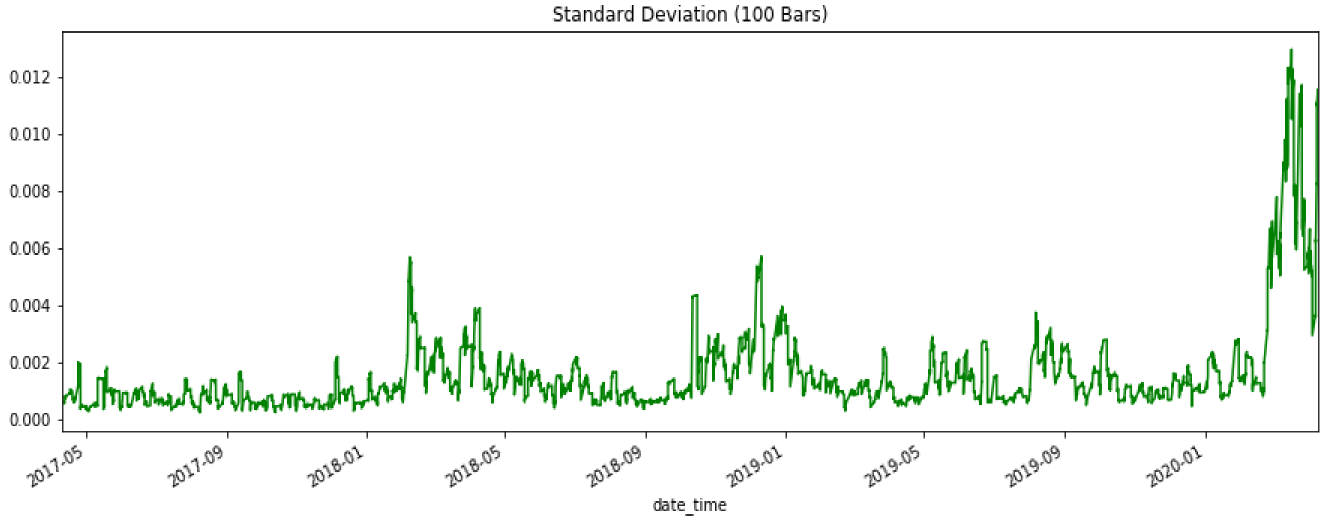


Figure C

CUSUM Filter Applied on a Geometric Brownian Motion

Threshold = 0.05

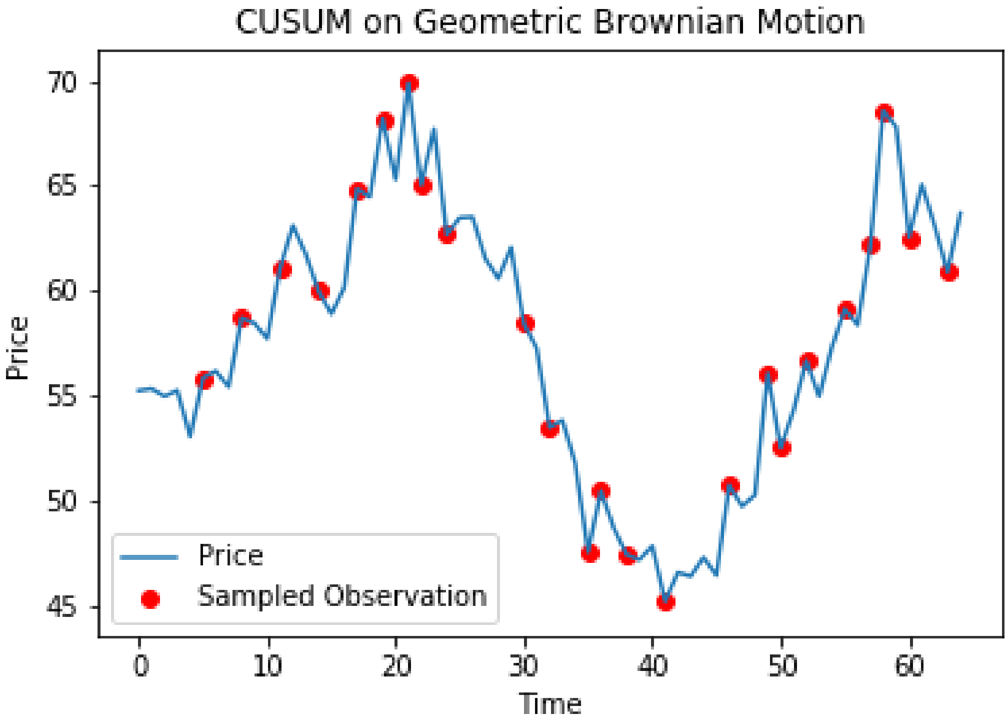


Figure D

Trend Scanning Applied to a Geometric Brownian Motion

Highly yellow points represent highly positive trends (determined via t-value) and Magenta represents highly negative.

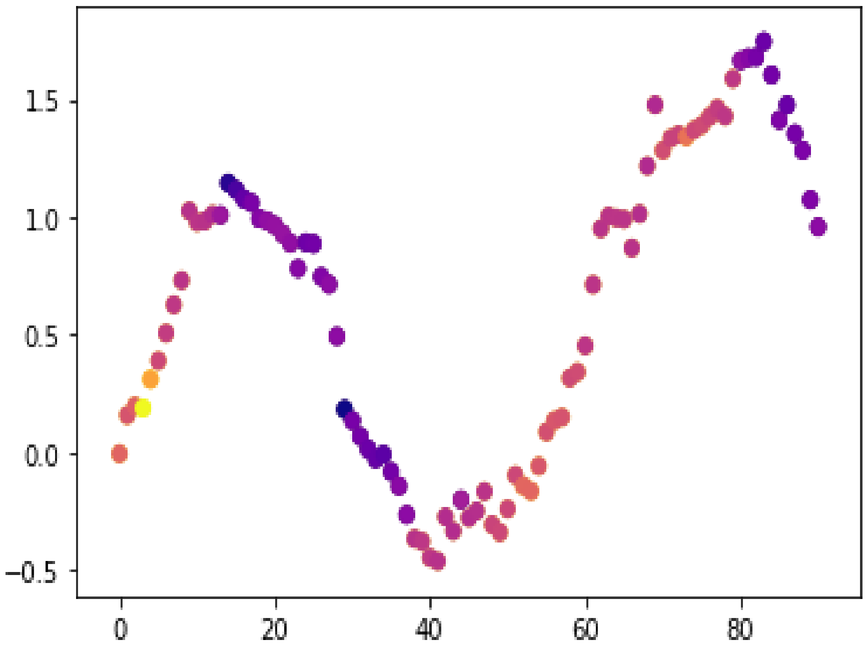


Figure E

HGY00 Plotted Against SPX: Signs of an impending crash

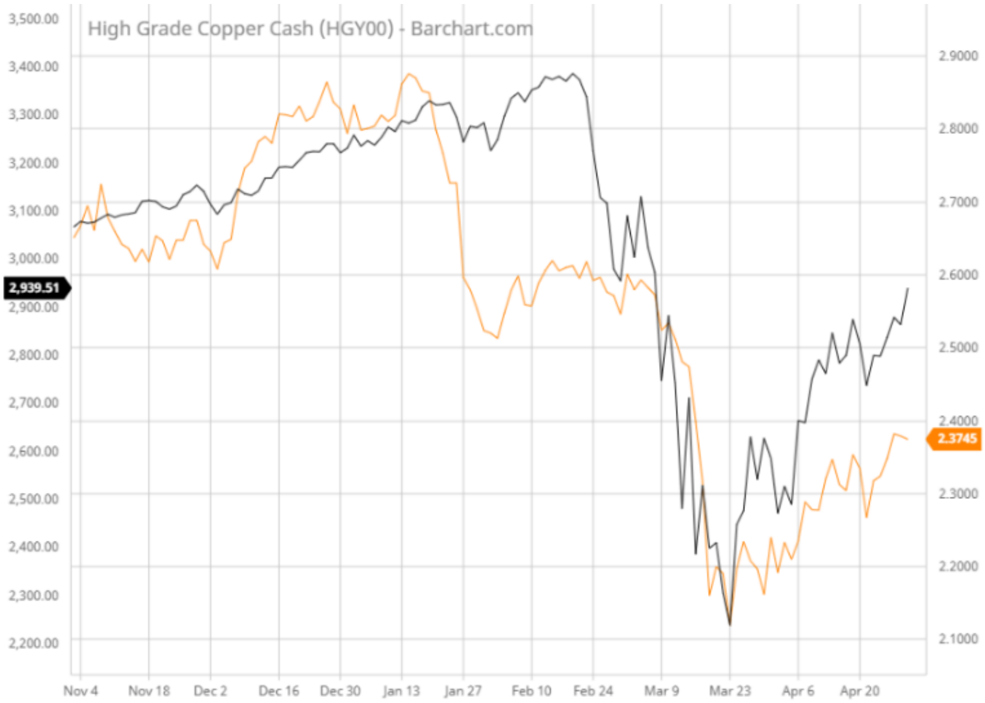
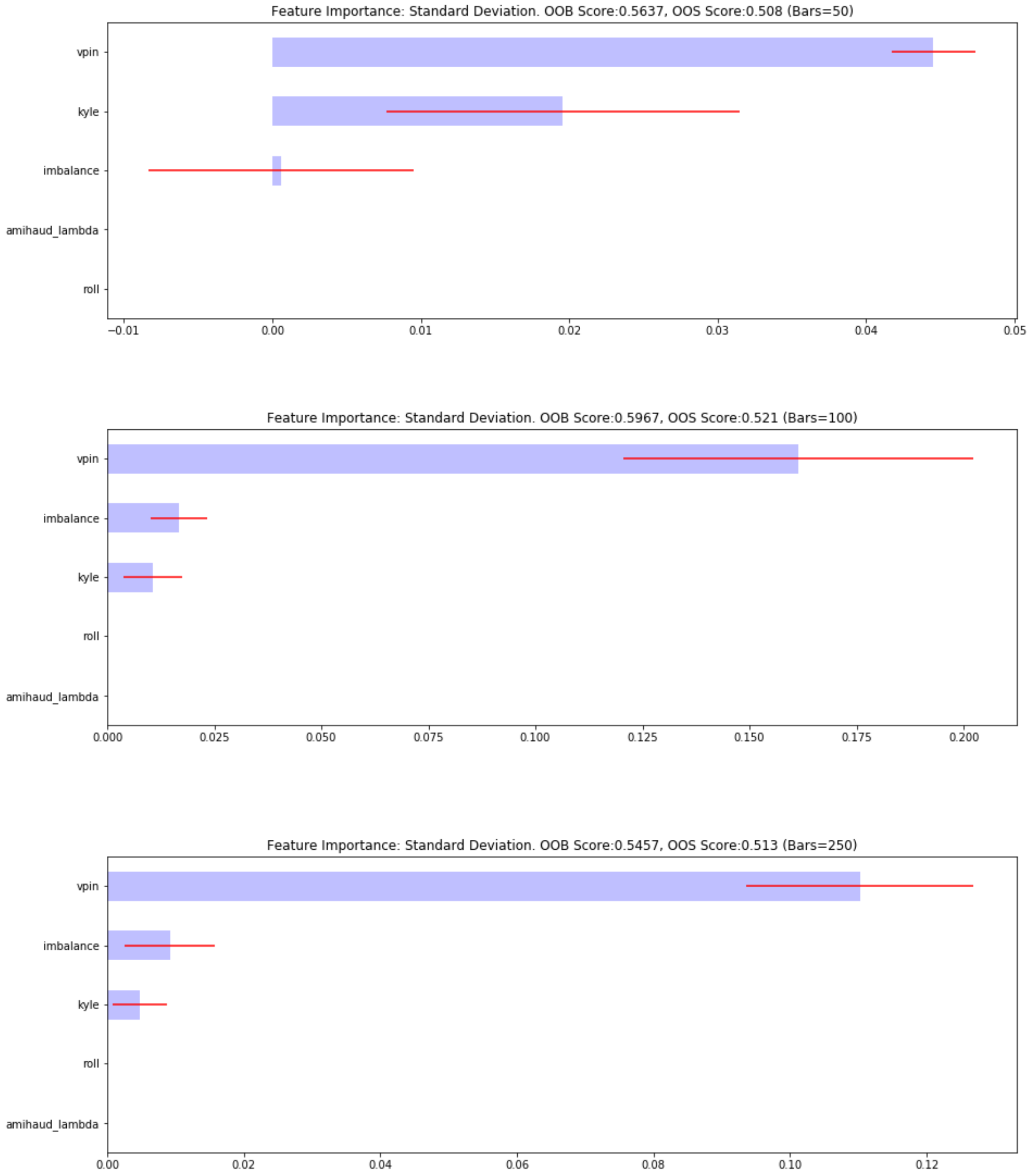


Figure F

Predicting Standard Deviation of Returns



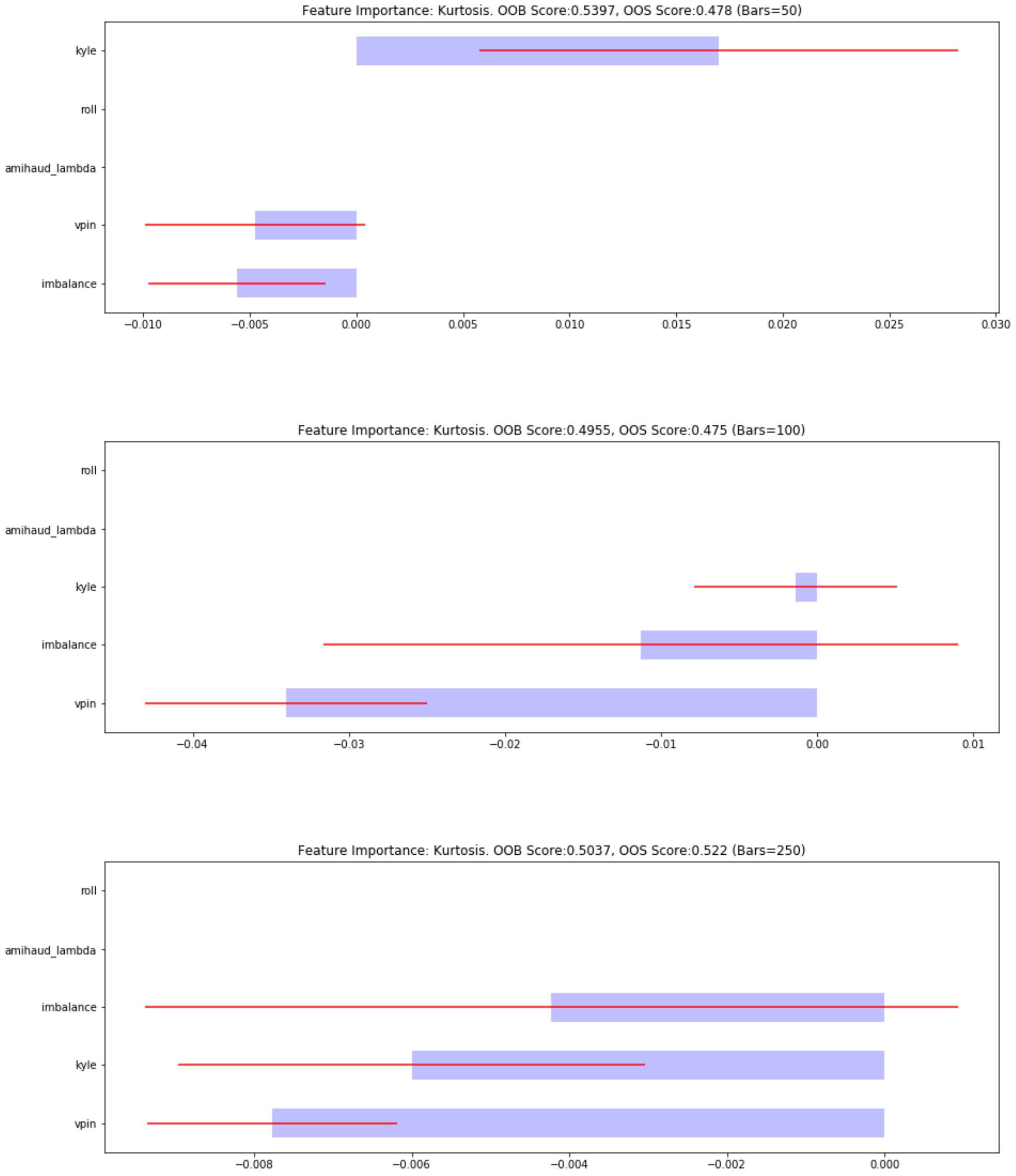
Feature Importance: Standard Deviation. OOB Score:0.5637, OOS Score:0.508 (Bars=50)

Feature Importance: Standard Deviation. OOB Score:0.5967, OOS Score:0.521 (Bars=100)

Feature Importance: Standard Deviation. OOB Score:0.5457, OOS Score:0.513 (Bars=250)

Figure G

Predicting Kurtosis



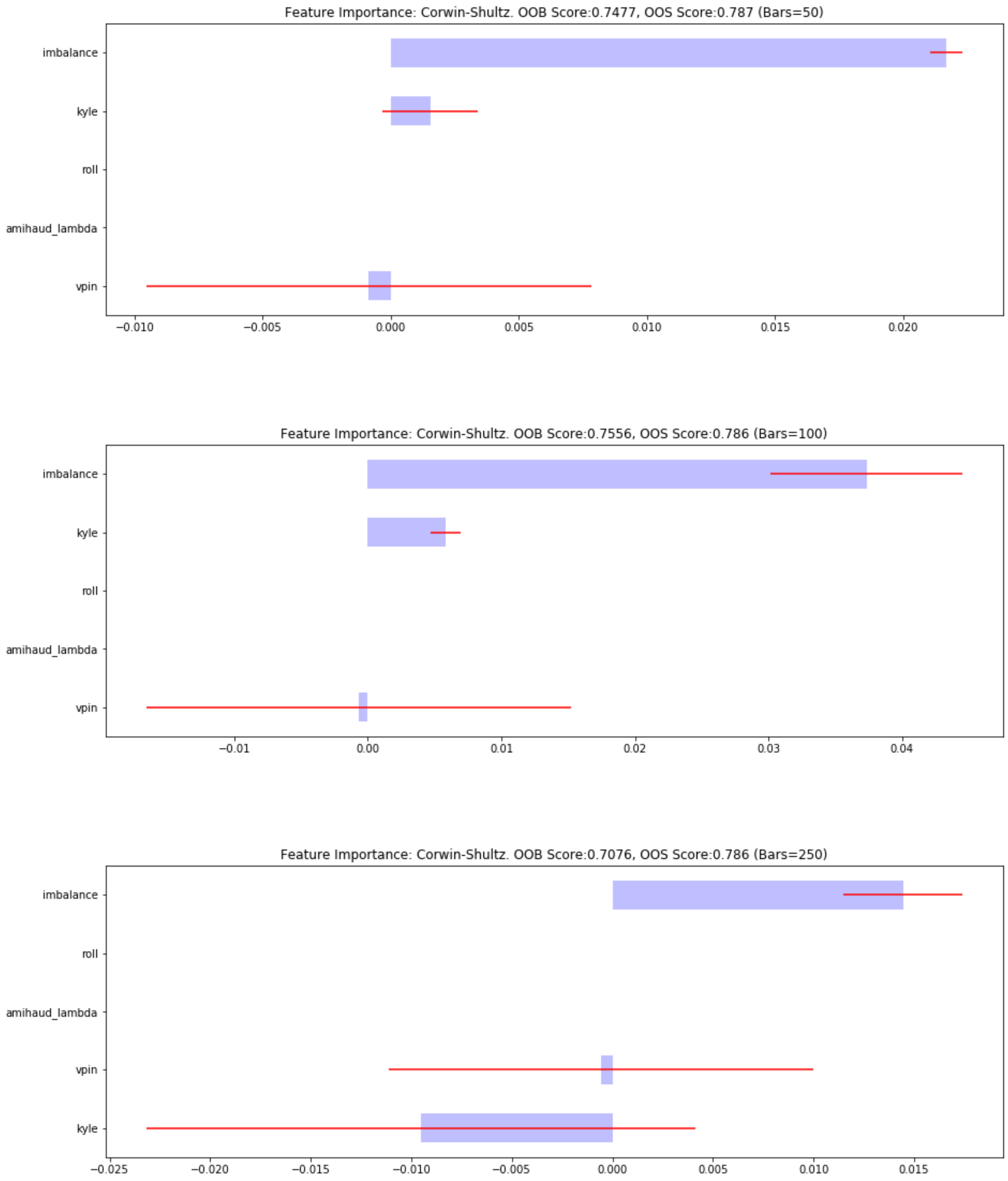
Feature Importance: Kurtosis. OOB Score:0.5397, OOS Score:0.478 (Bars=50)

Feature Importance: Kurtosis. OOB Score:0.4955, OOS Score:0.475 (Bars=100)

Feature Importance: Kurtosis. OOB Score:0.5037, OOS Score:0.522 (Bars=250)

Figure H

Predicting Corwin-Shultz


Feature Importance: Corwin-Shultz. OOB Score:0.7477, OOS Score:0.787 (Bars=50)

Feature Importance: Corwin-Shultz. OOB Score:0.7556, OOS Score:0.786 (Bars=100)

Feature Importance: Corwin-Shultz. OOB Score:0.7076, OOS Score:0.786 (Bars=250)

References

Amihud, Y., & Mendelson, H. (1987). Trading Mechanisms and Stock Returns: An Empirical Investigation. *The Journal of Finance*, *42*(3), 533–553. JSTOR. https://doi.org/10.2307/2328369

Bailey, D. H., & Lopez de Prado, M. (2014). *The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality* (SSRN Scholarly Paper ID 2460551). Social Science Research Network. https://doi.org/10.2139/ssrn.2460551

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, *57*, 289–300.

Benjamini, Y., & Liu, W. (1999). A Step-Down Multiple Hypothesis Testing Procedure that Controls the False Discovery Rate under Independence. *Journal of Statistical Planning and Inference*, *82*, 163–170.

Big Data, for better or worse: 90% of world's data generated over last two years. (n.d.). ScienceDaily. Retrieved May 1, 2020, from https://www.sciencedaily.com/releases/2013/05/130522085217.htm

Colquhoun, D. (n.d.). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, *1*(3), 140216. https://doi.org/10.1098/rsos.140216

Corwin, S. A., & Schultz, P. H. (2011). *A Simple Way to Estimate Bid-Ask Spreads from Daily High and Low Prices* (SSRN Scholarly Paper ID 1106193). Social Science Research Network. https://papers.ssrn.com/abstract=1106193

Easley, D., Kiefer, N. M., O'Hara, M., & Paperman, J. B. (1996). Liquidity, Information, and Infrequently Traded Stocks. *The Journal of Finance*, *51*(4), 1405–1436. JSTOR. https://doi.org/10.2307/2329399

Easley, D., Engle, R. F., O'Hara, M., & Wu, L. (2008). Time-Varying Arrival Rates of Informed and Uninformed Trades. *Journal of Financial Econometrics*, *6*(2), 171–207.

Easley, D., Lopez de Prado, M., & O'Hara, M. (2012a). *Flow Toxicity and Liquidity in a High Frequency World* (SSRN Scholarly Paper ID 1695596). Social Science Research Network. https://doi.org/10.2139/ssrn.1695596

Easley, D., Lopez de Prado, M., & O'Hara, M. (2012b). *The Volume Clock: Insights into the High Frequency Paradigm* (SSRN Scholarly Paper ID 2034858). Social Science Research Network. https://doi.org/10.2139/ssrn.2034858

Efron, B., & Hastie, T. (2016). *Computer Age Statistical InferenceL Algorithms, Evidence and Data Science* (1st ed.). Cambridge University Press.

Han, H., Guo, X., & Hua, Y. (n.d.). *Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest*. ResearchGate. http://dx.doi.org/10.1109/ICSESS.2016.7883053

Harvey, C. R. (2017). Presidential Address: The Scientific Outlook in Financial Economics. *The Journal of Finance*, *72*(4), 1399–1440. https://doi.org/10.1111/jofi.12530

Harvey, C. R., Liu, Y., & Zhu, C. (2015). *…and the Cross-Section of Expected Returns* (SSRN Scholarly Paper ID 2249314). Social Science Research Network. https://doi.org/10.2139/ssrn.2249314

Hur, J.-H., Ihm, S.-Y., & Park, Y.-H. (2017). *A Variable Impacts Measurement in Random Forest for Mobile Cloud Computing* [Research Article]. Wireless Communications and Mobile Computing; Hindawi. https://doi.org/10.1155/2017/6817627

IDC. (2014). *The Digital Universe of Opportunities: Rish Data and the Increasing Value of the Internet of Things*. https://www.emc.com/leadership/digital-universe/2014iview/index.htm

Kay, J. (2013). Enduring lessons from the legend of Rothschild's carrier pigeon. *The Financial Times.* https://www.ft.com/content/255b75e0-c77d-11e2-be27-00144feab7de

Kyle, A. S. (1985). Continuous Auctions and Insider Trading. *Econometrica*, *53*(6), 1315. https://doi.org/10.2307/1913210

Liquidity, Informatio. (n.d.).

Lopez de Prado, M. (2018). *Advances in Financial Machine Learning* (1 edition). Wiley.

Lopez de Prado, M. (2020a). *Machine Learning for Asset Managers*. Cambridge Elements.

Lopez de Prado, M. (2020b). *Clustered Feature Importance (Presentation Slides)* (SSRN Scholarly Paper ID 3517595). Social Science Research Network. https://doi.org/10.2139/ssrn.3517595

O'Hara, M. (1998). *Market Microstructure Theory* (1 edition). Wiley.

O'Hara, M. (2013). *High-frequency Trading*. Risk Books.

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance*, *39*, 1127–1139.

Romer, P. (2016). The Trouble with Macroeconomics. *The American Economist*.

Shen, D. (2015). *Order Imbalance Based Strategy in High Frequency Trading*. University of Oxford.

Solow, R. (2010). *Building a Science of Economics for the Real World: Statement prepared for the House Committee on Science and Technology*.

Toth, B., Palit, I., Lillo, F., & Farmer, J. D. (2015). Why is order flow so persistent? *Journal of Economic Dynamics and Control*, *51*, 218–239. https://doi.org/10.1016/j.jedc.2014.10.007